# USING MACHINE LEARNING TO PREDICT THE FUNCTINALITY OF WATER WELLS IN TANZANIA.

PRESENTER : AARON ONSERIO

CONTACT : aaron.onserio@student.moringaschool.com

GITHUB : https://github.com/AaronOnserio

# Business Overview

Challenges like extreme weather events due to climate change unprecedented population growth, forest clearance and land demarcations have all contributed to water crises in many parts of Africa. Many people in Africa particularly in Tanzania are experiencing water access, sanitation and hygiene crises.

Water is the basic essential need for human beings and yet more than half of the Tanzania population have no access to clean drinking water.

There are many water wells already established in Tanzania, yet some are completely not functioning, and need to be repaired.

# Business problem

- Tanzania has a population of 64.51 million and yet more than 40% of population have no access to clean drinking water. A holistic approach is needed to come up with a sustainable solution for the water wells in Tanzania.

# Objectives

- To Analyze the functionality of the water wells and find our which ones are working and which ones need repair.

- To build a classification model that predicts the functionality of water points.

# Data Understanding

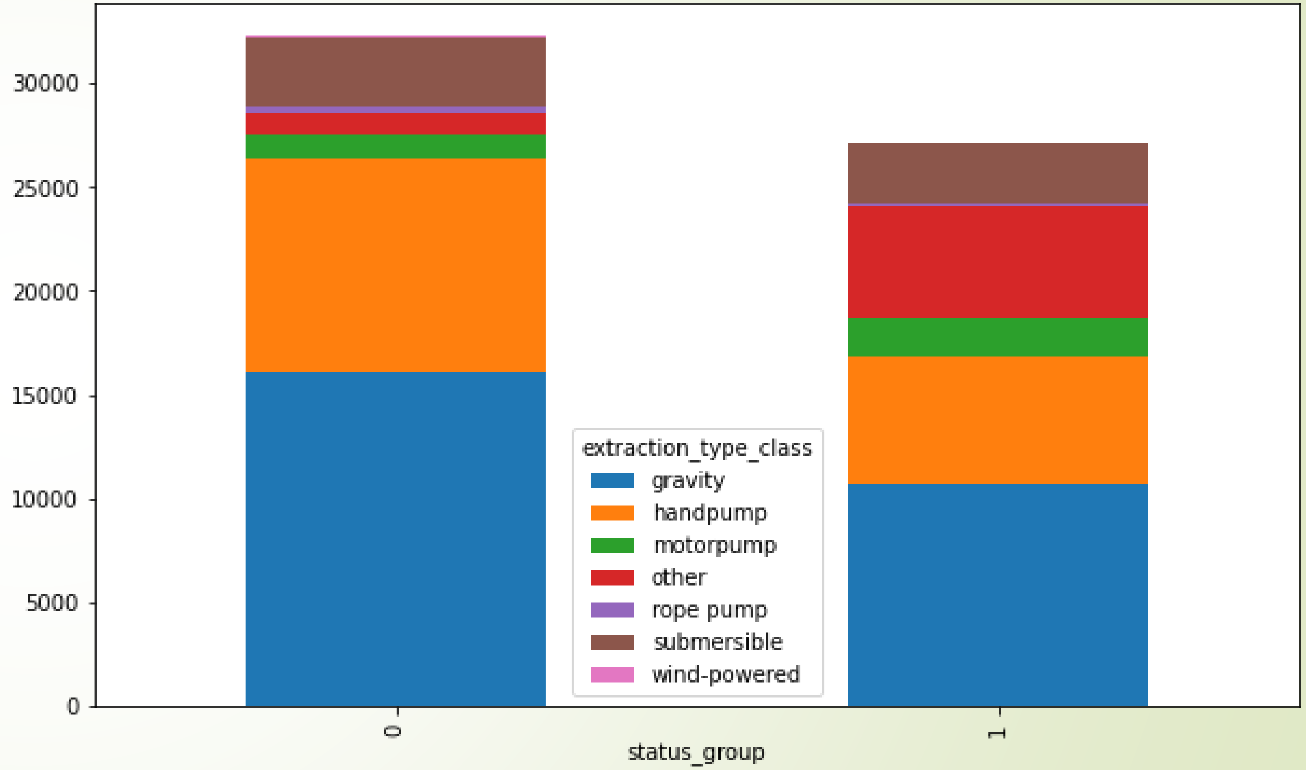
- The data used for this project is from the **Data Driven website**.
- The dataset contains nearly 60,000 entries rows and nearly 40 columns of water wells across Tanzania.
- Each record has information that includes various location data, technical specifications of the well, information about the water, etc.

# Methodology

This project focused on descriptive analysis, visualizations, and machine learning and modeling to describe trends for water wells in Tanzania that need of repair across the country.

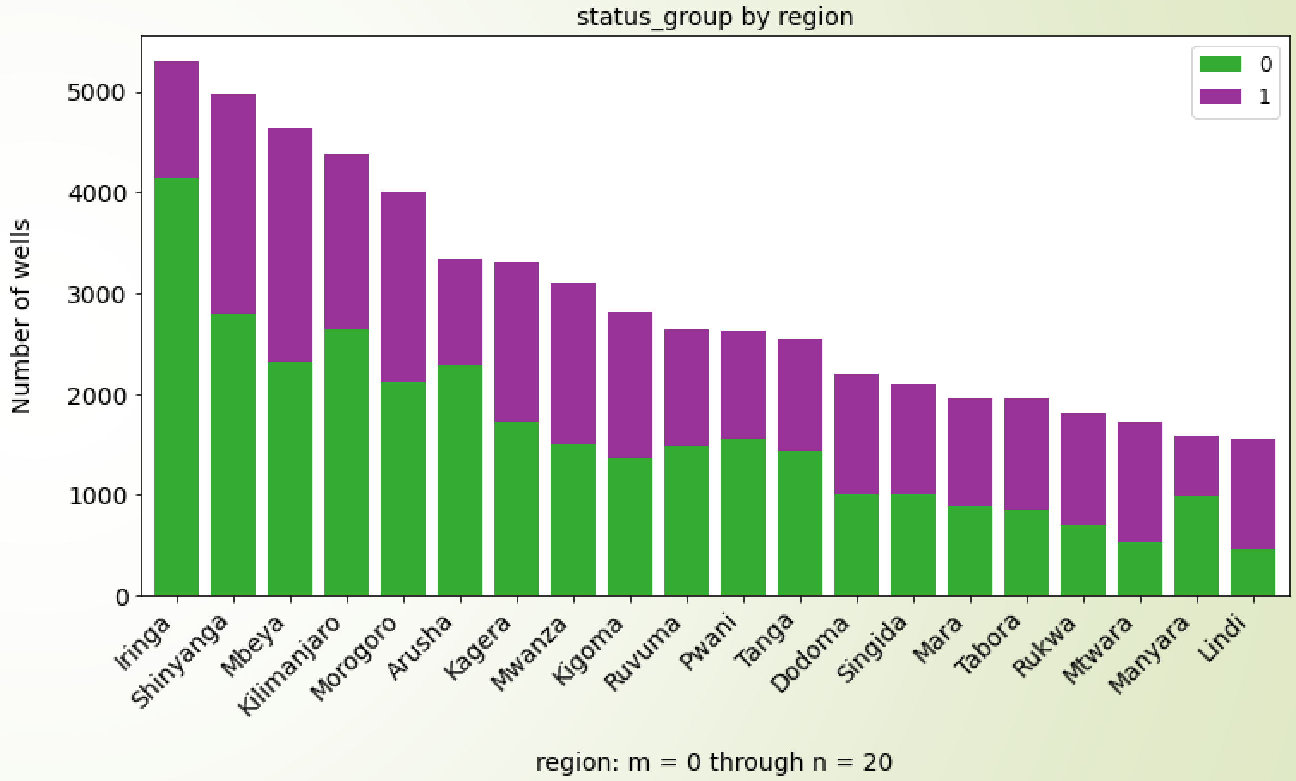
# Exploratory Data Analysis

**Extraction Type Class** - This graph shows the various methods people in Tanzania use to fetch the water from the wells at different places.



# Exploratory Data Analysis

**Region**- This displays various locations where we have wells in Tanzania

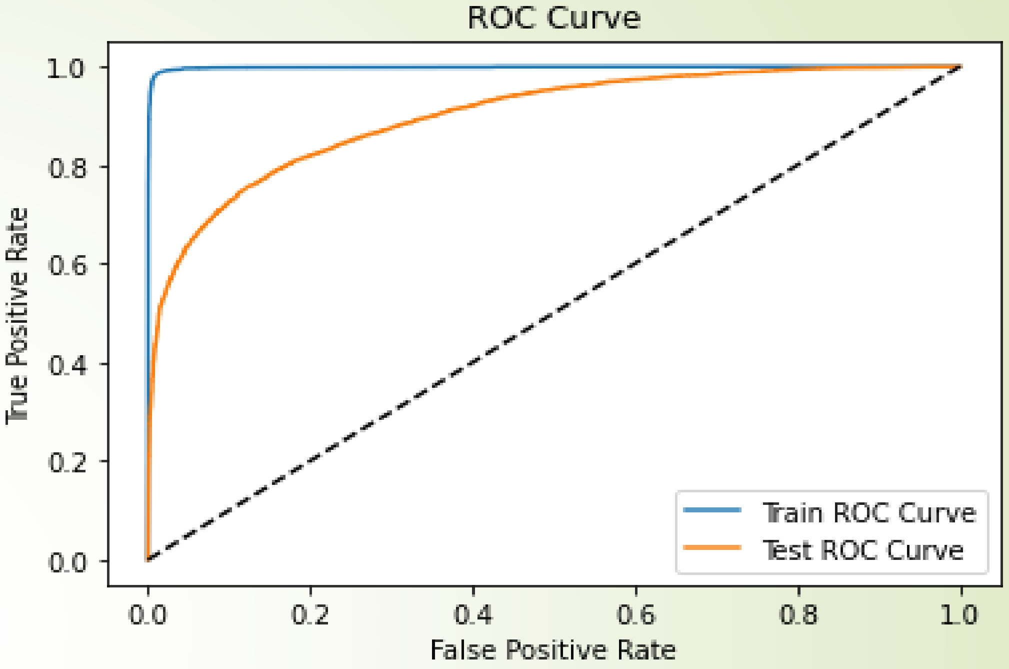


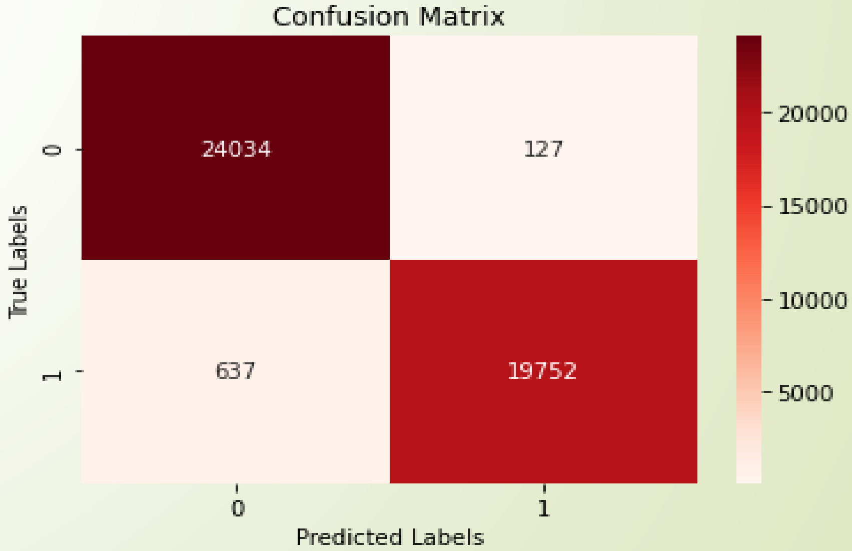
# Data Modeling

**Random Forest Classifier** - scores:
Train Accuracy: 0.98
Test Accuracy: 0.82
Train ROC-AUC: 0.99
Test ROC-AUC: 0.90

This model looks good The model was able to correctly predict accurately about 82% on test set and an ROC-AUC of 90% of the test set.

There was overfitting but given more time and more tuning it will be able to perform well.





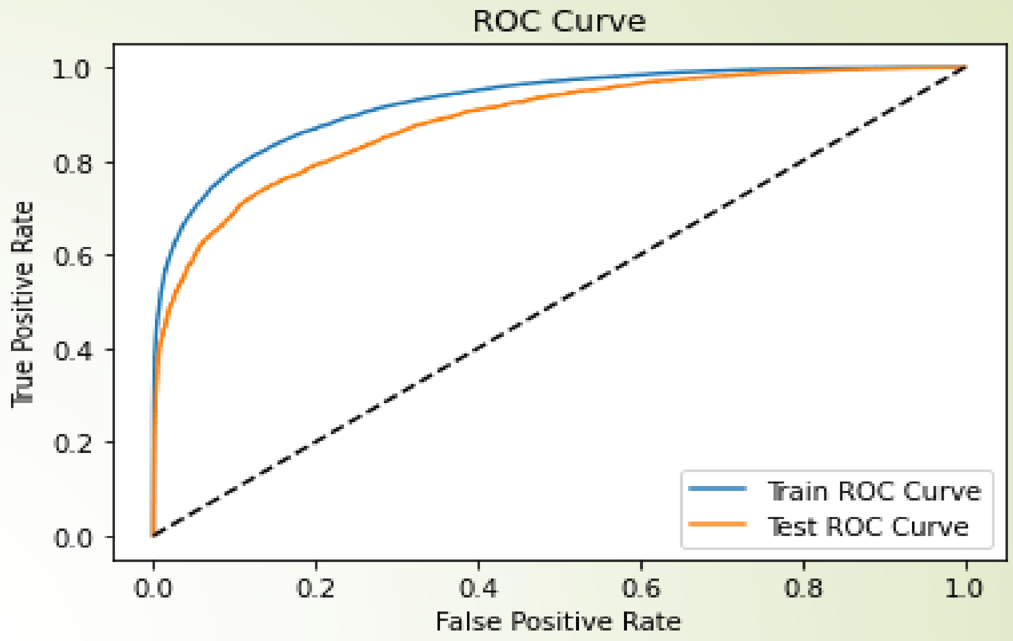
# Data Modeling

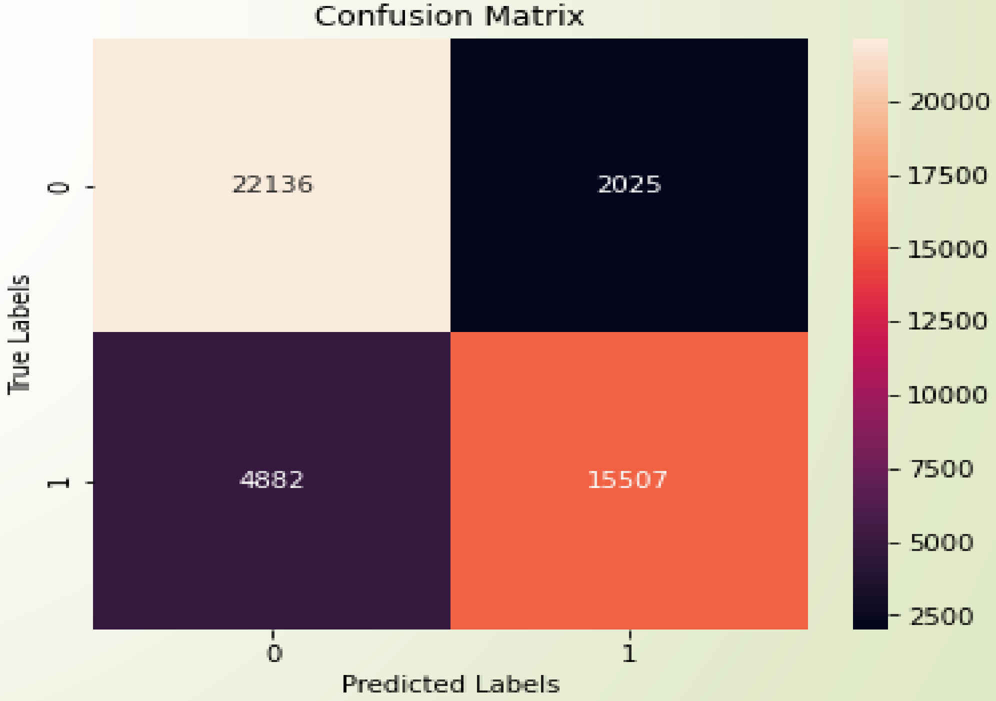
**XGBoost Classifier**

Train Accuracy: 0.84
Test Accuracy: 0.81
Train ROC-AUC: 0.93
Test ROC-AUC: 0.89

This model looks good The model was able to correctly predict accurately about 81% on test set and an ROC-AUC of 89% of the test set.

There was a slit overfitting but not that very significant. Given more time and more tuning it will be able to perform well.



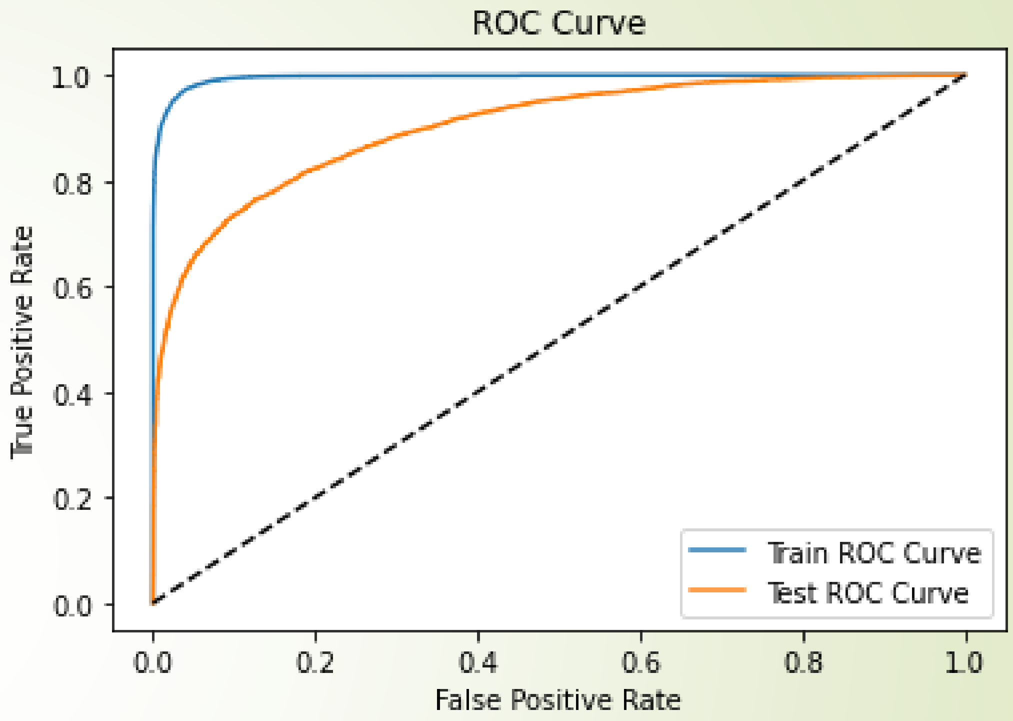


# Data Modeling

**Stack Classifier**
Train Accuracy: 0.96
Test Accuracy: 0.82
Train ROC-AUC: 0.99
Test ROC-AUC: 0.90

This model looks good The model was able to correctly predict accurately about 82% on test set and an ROC-AUC of 90% of the test set.

There was a slit overfitting but not that very significant. Given more time and more tuning it will be able to perform well.



# Results

- In this project I run several models to analyze and predict Tanzanian Water Wells that are functional, non-functional and those in need of repair.
- I considered several factors like accuracy, that measure how accurate the model can predict and ROC-AUC that measure how accurate the model identify the true positive and true negatives.
- Having said that The Stack Classifier Model has the best accuracy score of 82.19% on the test and ROC-AUC of 90.33% on the test dataset.

# Recommendations

- To get better results, especially on the wells that need repair class, more features need to be included during the data collection process. Adding information in the reporting of wells such as when the well was last serviced, what kind of repairs have been done on the wells, or if any parts have been replaced in the reporting of wells could prove to be useful for better predictions.

# Conclusion

- I first made a base model for each type of model classifier, trained and fitted with default parameters as a base.
- Thereafter, I selected key parameters to tune using sklearn GridSearchCV and the best parameters were used to run the final model.
- I compared the performance to the base model of each type, as well as between different model types.
- I evaluated the models using a classification report, a confusion matrix, ROC plots with AUC scores, and applied feature importance.

THANK YOU